Cow Rumen Metagenome Sequence Data Analysis Pipeline

Motivation. Cow rumen harbors a complex microbial community with thousands of species that efficiently degrade biomass into fermentable sugar, a bottleneck step in biofuel production. Discovering genes coding for cellulolytic enzymes and genomes harboring these enzymes are the scientific goals of the cow rumen metagenome analysis project. Initial cow rumen metagenome sequence datasets totaled 279 Gb of sequences consisting of 3 billion short reads with lengths from 75-125bp. Challenges to analyze these sequences include not only the size of the data and their suboptimal quality, but also the paucity of existing tools that can process data of this scale. To address these challenges a novel metagenome data processing has been developed.

Results. The metagenome data analysis pipeline consists of three components:

1. The **data preprocessing** component aims at improving the quality of metagenome sequence data. In the first step reads that are originated from spike-in controls, sequencing adaptors, sequencing artifacts, low complexity and tandem repetitive regions of the genome are removed. In the second step reads containing sequencing errors are trimmed using a k-mer based approach. Finally near identical reads originated from PCR amplification are grouped and collapsed into a consensus sequence by using Nary-tree and suffix tree data structures.
2. The **genome-based cellulase gene prediction** component. As illustrated in Figure 1 (left), the pipeline begins with selecting reads from abundant species by an empirical k-mer based approach, followed by assembling contig using Velvet and predicting gene using the IMG pipeline, and ends with predicting genes for cellulase signatures.

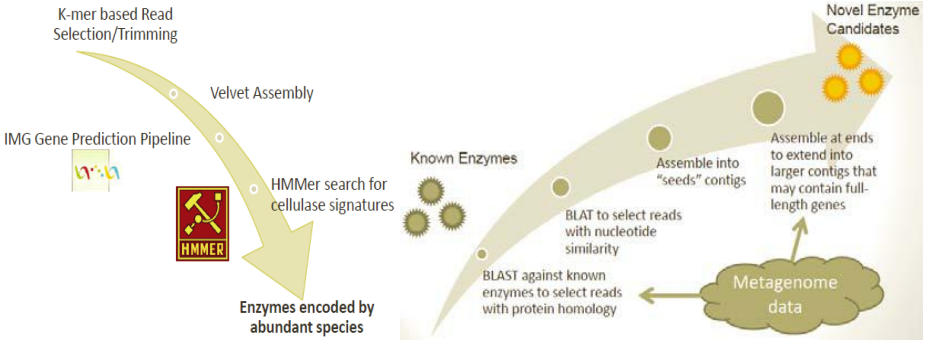

Figure 1. The generic enzyme prediction algorithms.

3. The **gene-based cellulase gene prediction**. As illustrated in Figure 1 (right), this part of the pipeline begins with selecting reads that are homologous to known enzyme coding genes, followed by assembling genes using Velvet and extending the resulted contigs, and ends with assigning novel candidates to known enzyme categories using BLAST.

The pipeline outlined above is limited to identifying enzymes that share significant similarity to the domains of existing known enzymes, and largely exclude novel enzymes that do not show homology but may possess novel enzymatic dynamics or novel substrate

The pipeline has been used successfully to predict novel enzyme candidates , with a significant portion of them expected to be true cellulases, as illustrated in Figure 2.

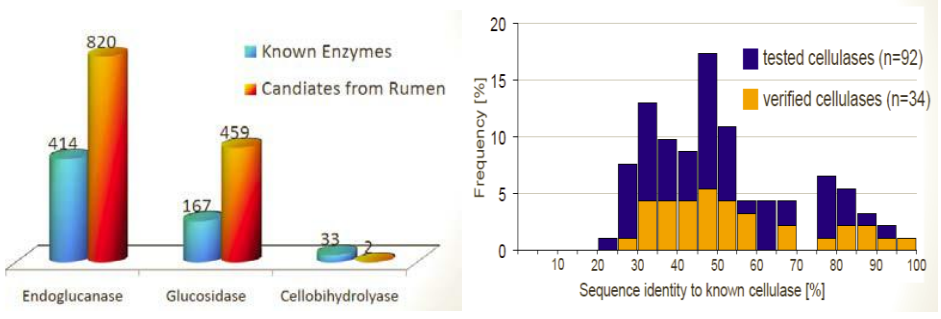

Figure 2. Many novel cellulase candidates are predicted from the cow rumen, and about 1/3 of the tested enzymes display biochemical activities *in vitro*.